Agentic AI-Based Smart Video Editor: A Multi-Agent Framework for Automated Video Summarization and Editing

Adarsh Poojary[1], Kunj Raut[2], Ayush Shetty[3], Tanmay Shigvan[4], Prof. Neha Nandurkar[5]
[1,2,3,4]*Student, Manjara Charitable Trust's Rajiv Gandhi Institute of Technology*
[5]*Professor, Manjara Charitable Trust's Rajiv Gandhi Institute of Technology*

***Abstract*—Automated video editing has emerged as a transformative application of artificial intelligence, addressing the growing demand for efficient content creation in the era of digital media. This study presents an agentic AI framework for intelligent video editing that integrates multi-agent architectures, vision-language models, and reinforcement learning techniques. We leverage pre-trained Vision-Language Models (VLMs) to extract semantic features from video content, enabling context-aware editing decisions. The system employs a multi-agent architecture where specialized agents collaborate to analyze video content, generate editing recommendations, and execute automated editing operations. Our framework incorporates reinforcement learning-based editing strategies to optimize sequential decision-making in video composition, moving beyond rule-based approaches to learn adaptive editing patterns from professional demonstrations. The implementation utilizes a Streamlit-based interface with MoviePy for video processing, demonstrating real-time analysis and editing capabilities. This integrated approach addresses both the technical challenges of automated video editing and the creative aspects of producing professional-quality content through intelligent agent collaboration and learned cinematographic principles.**

***Index Terms*—Agentic AI, Video Editing, Vision-Language Models, Multi-Agent Systems, Reinforcement Learning, Automated Content Creation, Deep Learning, Video Processing.**

I. INTRODUCTION

Early attempts at automation were often constrained to specific, event-driven scenes such as soccer game broadcasts or lecture recordings, relying on handcrafted features and empirical rules [3][7]. For instance, dedicated systems exist for generating highlight videos for baseball games using game logs, scoreboard recognition, and scene classification [7], or using a "core-around event model" to compose soccer highlights [3]. Such domain-specific systems use simpler computer vision algorithms to achieve rapid and accurate extraction of video segments [7]. More general editing tasks, which cover diverse scenes like movies or vlogs, have historically been rarely studied and remain non-trivial to automate effectively [5].

Recently, the field has accelerated with the integration of powerful modern artificial intelligence (AI) paradigms, especially Large Language Models (LLMs) and Deep Learning (DL) [4][6]. Deep learning models are now foundational for tasks like image and video caption generation, serving as the encoder-decoder architecture for turning visual data into descriptive language [4]. This linguistic and cognitive capability of LLMs is now being leveraged to create AI agents that act as collaborative assistants in the editing workflow [6].

This Agentic AI represents a significant advancement over traditional AI, allowing systems to undertake autonomous decision-making and adapt to complex, unstructured environments with reduced human intervention [10]. The core idea is to transform the editing process into a collaboration where the human user acts more as a director, using natural language (Language User Interface or LUI) to issue high-level commands, while the AI agent handles the complex execution and planning. The provided tool, a multi-agent framework for Smart Video Editing, directly embodies this trend by structuring the task as a sequence of specialized AI agent operations (Analysis, Editing, and Quality Control) to deliver on a user's descriptive editing prompt.

II. RELATED WORK

The complexity and labor-intensive nature of manual video editing, which requires multidisciplinary knowledge and skilled artists, motivated the development of semi- and fully automatic solutions [2]. Traditional editing systems were based on windowed user interfaces, supporting basic real-time operations like recording, playback, and cut, copy, and paste functions [1]. Early automatic systems focused on domain-specific scenes, such as live multi-camera directing for soccer matches or lectures [5][2].

Initial research focused on extracting highlights, employing simple visual cues or audio analysis [2]. Later systems evolved to use domain-specific logic, like a core-around event model for composing soccer highlights [3], or using game logs, scene classification, and scoreboard detection to rapidly extract events for baseball highlights [7]. Tools were also created for non-professional users to handle basic tasks like trimming and cropping online videos [8].

Despite these advances, two common research gaps limited the widespread adoption of automation [2][5]:

- Lack of Generality: Conventional methods relied on simple, hard-coded guidelines, limiting their flexibility and applicability to general scenes (e.g., vlogs or movies) [2][5]. Converting event-driven methods to a general context was often non-trivial [5].
- Aesthetic Gap: Automatically generated productions often lacked the artistic quality of human edits, resulting in an "aesthetic gap" due to the difficulty in transforming complex cinematic principles into computable formulas [5].

The advent of Deep Learning (DL) revolutionized this space by providing powerful feature extraction (CNNs) and language generation (RNNs/LSTMs), bridging the gap between visual information and meaningful description [4]. This foundation led to the current paradigm shift toward Agentic AI, which employs LLMs for planning, reasoning [10], and interpreting free-form language to execute complex editing goals [6][9]. This modern approach, epitomized by multi-agent frameworks, aims to overcome the "aesthetic gap" using advanced learning techniques [5], and the generality gap by designing agents capable of open-world task execution [9].

III. METHODOLOGY

The "Agentic AI-Based Smart Video Editor" utilizes a multi-agent framework to execute complex editing and summarization tasks based on high-level user commands. The methodology integrates Language User Interfaces (LUI) with modular AI components to achieve a highly autonomous Perception to Planning to Execution workflow [9][10].

1. Agentic Architecture and Input Processing

The system is built on an agent-based framework, embodying the modern trend toward autonomous intelligence [10].

- LUI/Perception: The process begins with a user's free-form language prompt (LUI) and an uploaded video. This command is interpreted by an LLM-powered agent to form a high-level editing goal [6].
- Decomposition/Planning: A master agent converts the complex goal into an action plan comprising sequenced sub-tasks (e.g., "summarize," "highlight," "trim") [6].

2. Core Video Analysis and Feature Extraction

Accurate editing requires robust content understanding, typically a bottleneck in traditional systems [2].

- Deep Feature Extraction: Deep Learning models, particularly Convolutional Neural Networks (CNNs), are used to encode visual content, extracting features and generating captions to map visuals to language [4]. This forms the basis of Visual Narrations for each clip, critical for semantic understanding [6].
- General Contextual Representation: To overcome scene-specificity, a pre-trained Vision-Language Model (VLM) is used to extract generalized editing-relevant attribute distributions of content shots, providing the context for subsequent decisions [5].

3. Multi-Agent Execution and Optimization

The core editing is performed sequentially by specialized agents.

- Analysis Agent: This agent extracts metadata and identifies key moments in the video, feeding a content map to the subsequent editing layer.
- Editing Agent: The agent selects the appropriate video manipulation (e.g., summarization,

highlighting, trimming) based on the classified user prompt, executes the edit using modular operations, and manages clip sequencing (storyboarding) [6].

- Refinement and Quality Control: An optional Quality Control Agent performs final optimization and compression. For advanced stylistic tasks (e.g., surpassing the aesthetic gap), Reinforcement Learning (RL) is employed to train a virtual editor for superior sequential editing decisions, learning patterns from professional demonstrations [5].

4. Technology Stack

The system is prototyped using Python with frameworks such as Streamlit, Gemini API, and dedicated video processing libraries (e.g., moviepy. editor) to demonstrate functional agent workflows.

IV. RESULT AND ANALYSIS

The implementation of Agentic AI-based Smart Video Editors demonstrates significant success in automating complex and labor-intensive editing tasks, especially those related to video summarization and content structuring [2][6]. Current systems showcase strong performance across core video editing challenges:

- High Accuracy in Domain-Specific Extraction: For structured content like sports videos, specialized systems achieve high accuracy, such as 98.18% in extracting trimmed event videos for baseball games [7]. This is accomplished by leveraging game logs and rule-based scene classification [7].
- Effective Content-Aware Operations: Even in systems providing simple utilities, focusing on efficiency (like performing Trim > Crop > Rotate > Disable Audio) significantly speeds up video encoding compared to other sequential combinations [8].
- Successful Linguistic-to-Visual Task Execution: Modern LLM-based agents, which interpret user commands and video content during a perception > planning > execution workflow, demonstrate competitive results in tasks like video salient object detection and open-world reframing [9]. The ability of LLMs to analyze extracted visual narrations allows for semantic-based video retrieval and intelligent storyboarding [6].

Analysis reveals three key areas of impact and remaining challenges:

- Bridging the Generality Gap: The use of Vision-Language Models (VLM) is crucial, as their knowledge is transferred to extract editing-relevant representations that generalize across diverse scenes, moving beyond the traditional reliance on domain-specific handcrafted features [5]. This shift supports the goal of creating editors for general-purpose content like movies and vlogs [5].
- Addressing the Aesthetic Gap with RL: Despite high accuracy in technical tasks, a gap remains between automated output and the nuanced style of professional editing [5]. To overcome this, Reinforcement Learning (RL) frameworks are being introduced to train virtual editors for sequential decision-making, which helps optimize video quality and learn professional editing patterns from demonstrations [5].
- Ensuring User Agency and Trust: The introduction of LLM-powered agents necessitates a focus on user experience [6]. The "plan-and-execute" architecture, where the agent proposes a plan based on the user's free-form prompt before execution, allows the user to approve or revise actions, ensuring they maintain agency and control over the final creative outcome [6][10]. This dual modality of agent assistance and direct manipulation helps preserve the user's artistic vision [6].

Overall, Agentic AI, with its capacity for autonomous decision-making and adaptive learning, is rapidly transforming video editing from a technical manipulation task into a collaborative process of co-creation [10][6].

V. CONCLUSION

This paper presents a cohesive framework for an Agentic AI-Based Smart Video Editor that uses a multi-agent architecture to transform video editing into an intelligent, collaborative process. The system addresses the traditional challenges of generality and quality in automated production. It achieves this by leveraging LLMs and Vision-Language Models (VLMs) to interpret high-level Natural Language Instructions (LUI), converting abstract creative goals into actionable plans via a perception > planning > execution pipeline. The core architectural innovation

involves orchestrating specialized agents (Analysis, Editing, and Quality Control) for sequential tasks, enabling complex functions such as semantic-based content analysis, intelligent summarization of "key moments," and dynamic reframing. To overcome the "aesthetic gap," the framework incorporates Reinforcement Learning (RL), allowing the editor to learn professional editing styles and make superior sequential decisions. This synthesis paves the way for a new generation of trustworthy co-creation tools by combining autonomous decision-making with robust user control.

REFERENCES

[1] P. Vankat Rangan, H. M. Vin, K. Chan and I. A. Aaberg, "A window-based editor for digital video and audio," Proceedings of 25th Hawaii International Conference on System Sciences, Kauai, HI, USA, 1992, pp. 640-648 vol.2, doi: 10.1109/HICSS.1992.183315.

[2] Xinrong Zhang, Yanghao Li, Yuxing Han, and Jiangtao Wen, "AI Video Editing: a Survey.", 2022.

[3] Kai Jiang, Xiaowu Chen and Qinping Zhao, "Automatic composing soccer video highlights with core-around event model.", in 2011 12th International Conference on Computer-Aided Design and Computer Graphics, 2011, pp. 183–190.

[4] Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap.", IEEE Access, vol. 8, pp. 218386–218400, 2020.

[5] Panwen Hu, Nan Xiao, Feifei Li, Yongquan Chen, and Rui Huang, "A Reinforcement Learning-Based Automatic Video Editing Method Using Pre-trained Vision-Language Model," in Proc. 31st ACM Int. Conf. Multimedia (MM '23), Ottawa, ON, Canada, pp. 1–10, Oct. 2023.

[6] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi, "LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing.", in 29th International Conference on Intelligent User Interfaces (IUI '24), 2024.

[7] Younghyun Lee, Hyunjo Jung, Cheoljong Yang, and Joonsoo Lee, "Highlight-Video Generation System for Baseball Games.", ICCE-Asia, Seoul, Korea (South), 2020, pp. 1-4, doi: 10.1109/ICCE-Asia49877.2020.9277391.

[8] G. Vasanth, J. Heilman, P. Shetty, H. Amin and M. Geetha, "VideoCutTool - Online Video Editor Tool for Wikimedia Commons," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0823-0827, doi: 10.1109/ICCSP48568.2020.9182122.

[9] Jiawang Cao, Yongliang Wu, Weiheng Chi, Wenbo Zhu, Ziyue Su, and Jay Wu, "Reframe Anything: LLM Agent for Open World Video Reframing.", arXiv preprint arXiv:2403.06070, 2024.

[10] Siva Kumar Mamillapalli, "The Agentic AI Framework: Enabling Autonomous Intelligence.", IJIRMPS, vol. 13, no. 1, pp. 1–4, 2025.